Földtani Közlöny 136/1, 139–146. (2006) Budapest

A minta értelmezési problémái: elmélet és gyakorlat

Understanding of spatiotemporal samples: a practical view for geologists

KOVÁCS József[1] – KOVÁCSNÉ SZÉKELY Ilona[2]

(5 ábra, 4 táblázat)

Tárgyszavak: minta, környezetszennyezés; statisztikák
Keywords: spatiotemporal sample, environmental pollution; statistics

Abstract

One of the most fundamental concept of statistics is the (random) sample. Our experience – acquired during the years of undergraduate education – showed that prior to industrial practice, the students in geology (a most probably in many other non-mathematics oriented disciplines as well) are often confused by the possible multiple interpretation of the sample. The confusion increases even further, when samples from stationary temporal, spatial or spatio-temporal phenomena are considered. Our goal in the present paper is to give a viable alternative to this overly mathematical approach, which is proven to be far too demanding for geological students.

Using the results of an environmental pollution analysis we tried to show the notion of the spatiotemporal sample and some of its basic characteristics. On the basis of these considerations we give the definition of the spatiotemporal sample in order to be satisfactory from both the theoretical and the practical points of view.

Összefoglalás

A statisztika alapfogalmai között talán a legalapvetőbb a minta. A gyakorlat és az azt megelőző felsőoktatási tapasztalatok azt bizonyítják, hogy a minta értelmezése nehézséget jelent. Ennek oka, hogy az adatok elemzése során a szakemberek a mintából számított alapstatisztikaként (átlag, szórás és egyéb mutatók) egy-egy számértékkel dolgoznak és nem merül fel a munka során, hogy ez egy valószínűségi változó, ugyanis a minta realizációi mintáról mintára változnak, így a belőlük számított statisztikák is változni fognak. A publikáció egy környezetszennyezés elemzési eredményeinek felhasználásával megpróbálja bemutatni a minta fogalmát.

Bevezetés

A statisztika alapfogalmainak és módszereinek megértése és helyes alkalmazása elképzelhetetlen valószínűségszámítási alapok nélkül. Ilyen és talán legalapvetőbb fogalom a minta (NEMETZ & WINTSCHE 1999, ANDERSON & LOYNES 1987).

Az ipari gyakorlatban a „minta" fogalma nem teljes körűen és nem egyértelműen definiált. Sok esetben ugyanis a matematikai értelemben vett „minta" egy elemét tekintik mintának, de ugyanígy „minta" névvel illetik az azonos helyen és időben, azonos paraméterre végzett elemzések átlagát, vagy például a földtanban valamely X,Y koordinátával azonosítható helyen a minőségi paraméterek vastagsággal súlyozott átlagát is (FÜST 1998, WEBSTER & OLIVER 1990).

[1]Eötvös Loránd Tudományegyetem, TTK, Alkalmazott és Környezetföldtani Tanszék, 1116, Budapest, Pázmány Péter sétány 1/c.
[2]Budapesti Gazdasági Főiskola, KVIFK, Módszertani Intézet, 1054 Budapest, Alkotmány ut 9–11.

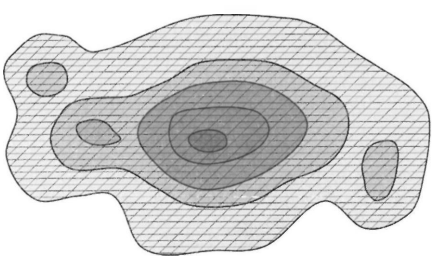
Az ipari gyakorlatot megelőző felsőoktatási tapasztalatok is azt bizonyítják, hogy a minta többféle értelmezése mind a matematikus, mind a geológus hallgatóknak nehézséget jelent. Ennek az az oka, hogy amikor a hallgató egy szakmai, például geokémiai feladatot lát, a mintának egy realizációját kapja és ezt azonosnak tekinti a mintával. A mintából számított alapstatisztikaként (átlag, szórás és egyéb mutatók) egy-egy számértékkel dolgozik, és nem érti azon állítást, hogy ez egy valószínűségi változó, ugyanis a minta realizációi mintáról mintára változnak, így a belőlük számított statisztikák is változi fognak.

A minta elméleti fogalma

A statisztikai minta az X valószínűségi változóra vonatkozó véges számú független megfigyelés eredménye $X=(X_1,X_2, \ldots ,X_n)$, ahol $X_1,X_2, \ldots ,X_n$ egymástól független, azonos eloszlású valószínűségi változók. A minta elemeinek eloszlása megegyezik a sokaság eloszlásával és $E(X_i)= m$; $D(X_i)= \sigma$ ahol i=1, 2, ... n.

A minta realizációja a megfigyelések számszerűsített értékei $(x_1, x_2,\ldots,x_n)$, ha egy konkrét mintavételnél $X_1=x_1$, $X_2=x_2$, ... , $X_n=x_n$ adódik.

Ez a megfogalmazás nehezen érthető, különösen nem matematikusok számára. Induljunk ki ezért egy másfajta megközelítésből. Belátható, hogy valamely időben és térben változó természeti jelenség adott időponthoz rendelhető három dimenziós metszete, elméletileg végtelen számú (N= ∞), „nulla térfogatú" (V=0) elemi részre osztható. Ebből következően a jelenség kutatása során valamely változó (földtani szóhasználatban paraméter) vonatkozásában egyetlen olyan minta realizáció állítható elő, melynek elemszáma végtelen. Elméletileg ez az adathalmaz tekinthető a vizsgált sokaságnak. Ennek elméletileg meghatározható a várható értéke (m) és a szórása (σ). Amennyiben az elemi részek „térfogata" továbbra is nulla, de N< ∞, a minta realizációk száma (n) növekszik. Minden realizációból meghatározhatóak a statisztikai jellemzok (például: mintaátlag ($\bar{X}$) mintaátlag szórása ($\sigma_{\bar{X}}$), egyes megfigyelések szórása (s). A végtelen sokaság szemléltetéséhez tekintsük meg az *1. ábrát.*

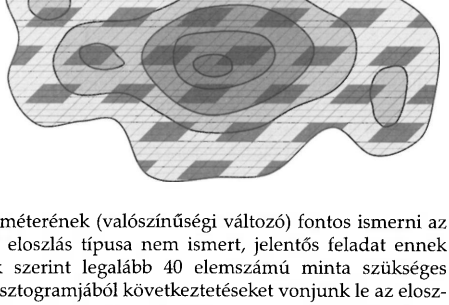


1. ábra. Egy elméleti jelenség egy adott paraméterének izovonalas képe. A jelenség „területét" végtelen számú „nulla térfogatú" elemi részre osztottuk fel

Fig. 1 Isoline figure of a "parameter" of a theoretical phenomenon

Valamely jelenség kutatása során – főként anyagi okok miatt meglehetősen ritkán adódik arra lehetőség, hogy olyan (esetenként több ezer) elemszámú mintát vegyünk, amelyből a statisztikai jellemzők nagy pontossággal számíthatók.

A gyakorlatban egy másik problémával is szemben találjuk magunkat. Ez pedig az, hogy a minta elemi részeinek „térfogata" nem nulla, hanem nullánál nagyobb és mindenképpen mérhető nagyságú. Az esetek zömében nincs arra lehetőség, hogy közel végtelen elemszámú mintát vegyünk, vagy a mintavételt kisebb elemszám mellett többször megismételjük. A helyzet tehát az, hogy $V>>0$, de a jelenség egészéhez képest $V \approx 0$, azaz a minta elemek térfogata és a szórás közötti kapcsolatot figyelmen kívül lehet hagyni, de $N<<\infty$. Ekkor felmerül a kérdés, hogy vajon az ilyen minta reprezentatívnak tekinthető-e, azaz a minta valóban híven tükrözi a sokaságot, amelyből származik. A 2. *ábra* egy lehetséges minta realizációt mutat $N<\infty$ és $V>0$ esetére.



2. *ábra.* Egy lehetséges minta realizáció, $N<\infty$ és $V>0$ esetén

Fig. 2 A possible sample realisation for $N<$ and $V>0$

Egy jelenség valamely paraméterének (valószínűségi változó) fontos ismerni az eloszlását. A gyakorlatban az eloszlás típusa nem ismert, jelentős feladat ennek meghatározása. Tapasztalatok szerint legalább 40 elemszámú minta szükséges ahhoz, hogy egy paraméter hisztogramjából következtetéseket vonjunk le az eloszlás típusára. A minta alapján becslés adható az eloszlás legfontosabb jellemzőire, a várható értékre és a szórásra.

Tekintsük az $(X_1, X_2, \ldots, X_n)$ minta átlagát $\overline{X} = \frac{\sum X_i}{n}$

A mintaátlag jól közelíti a várható értéket mivel $E(\overline{X})=m$ és a minta elemszámának növelésével a mintaátlag szórása csökken a következők szerint

$$D(X) = \frac{\sigma}{\sqrt{n}}$$

A σ szórás az úgynevezett korrigált tapasztalati szórással becsülhető az alábbi összefüggés alapján:

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$

A matematikai statisztika eszköztárában van lehetőség arra, hogy az adott minta alapján megadható legyen egy intervallum, amely előre rögzített valószínűséggel tartalmazza a várható éréket, ha a minta egy normális eloszlású sokaságból származik. Ez az intervallum megadható akkor is, ha a sokaság nem normális, de ismert

eloszlású és a minta elemszáma elég nagy. Ezt a megvalósítást a centrális határeloszlás tétele biztosítja (lásd később). Természetesen, mintáról mintára az intervallum helyzete is változik. Ha a minta elemszámát növeljük egyre szűkebb lesz az intervallum, amelyben a várható érték adott valószínűségi szinten elhelyezkedik. Ha tehát előre megadjuk, hogy adott valószínűségi szinten mekkora lehet annak az intervallumnak a szélessége (más szavakkal, a megengedett hiba), amelybe a várható érték beleesik, ebből a minta szükséges elemszáma kiszámítható. Az előbb említett intervallum szélessége (Δ) kifejezhető a normális eloszlás vagy a t-eloszlás táblázatában található, adott valószínűségi szinthez tartozó kritikus értékének és a mintaátlag szórásának szorzataként:

$$\Delta = z \cdot \frac{\sigma}{\sqrt{n}} \quad \text{vagy} \quad \Delta = t \cdot \frac{s}{\sqrt{n}}$$

A fenti összefüggésekből a Δ hibához szükséges mintaelemszám meghatározható:

$$n = \left(\frac{z \cdot \sigma}{\Delta} \right)^2 \quad \text{vagy} \quad n = \left(\frac{t \cdot s}{\Delta} \right)^2$$

Bizonyos esetekben a minta elemszámának ismerete nem elegendő és további feltételeknek kell teljesülnie. Például ásványlelőhelyek kutatásánál abból a feltételezésből indulunk ki, hogy az ásványlelőhely mint természeti képződmény folytonos tulajdonságú paraméterekkel rendelkezik. Ez a feltételezés csak részben igaz (gondoljunk például a tektonikára, vagy az aranybányászatban a röghatás jelenségére), ugyanakkor ezt a feltételezetten folytonos tulajdonságot olyan minta alapján próbáljuk megismerni, amelynek elemei diszkrétek. A folytonossági alapfeltétel továbbvitele, amikor a diszkrét mintaelemek hisztogramjának alakjából valamely folytonos eloszlással való közelítés lehetőségét tételezzük fel.

Kérdésként merülhet fel, hogy a mintaátlag – mint valószínűségi változó – milyen eloszlást követ, függetlenül attól, hogy milyen volt a sokaság eloszlása, amelyből a minta származott. A választ a centrális határeloszlás tételének egy alkalmazása adja meg.

A centrális határeloszlás tétele azt mondja ki, hogy ha X_1, X_2, ..., X_n azonos eloszlású, független és véges szórású valószínűségi változók és $E(X_i)=m$, $D(X_i)=\sigma$, $(i=1, 2,..., n)$ akkor a változók összege $\sum_{i=1}^{n} X_i$ közel normális eloszlást követ, ha az n elég nagy.

Így a mintából számított $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ mintaátlag közel normális eloszlású valószínűségi változó, ha az n mintanagyság elég nagy. Ennek a bizonyítása jól ismert, ezért ettől itt eltekintünk és helyette, egy gyakorlati megközelítést választunk.

A minta fogalma és néhány tulajdonsága egy példa tükrében

A gyakorlatban szinte megoldhatatlan anyagi nehézségekbe ütközik olyan nagy elemszámú minta előállítása, ami be tudja tölteni a vizsgált sokaság szerepét. Ezért jelentős előrelépés, ha egy valós környezetszennyezés mért kémiai komponen-

seinek adathalmazát fel tudjuk használni a minta fogalmának tisztázására, paramétereinek kiszámítására és eloszlásának meghatározására.

Az 1900-as évek elejétől, Budapest XXII. kerületének egy részén, Nagytétény sűrűn lakott területén, egy gyárüzem szennyező forrásként üzemelt, kéménye káros anyagokat bocsátott ki. A lakóingatlanokat jelentős nehézfém terhelés érte, amely a területhasználattal a talajszelvény nagyobb mélységeibe is bekeveredett. A 3,5 km^2-en fekvő ingatlanok környezetvédelmi feltárásait 20 cm-enként mintázták. Esetünkben a felszíntől számított 0–20 cm-es szintben több mint 1000 megfigyelés történt, aminek során 24 elemre vonatkozó elemtartalom vizsgálatát végezték el. Ezekből három elemet választottunk ki: kalcium, foszfor, arzén, amelyek a statisztikai modellben valószínűségi változók. A mértékegység mindegyik esetben mg/kg (ANDÓ & BATA 2001; BATA & ANDÓ 2005; BATA et al. in press).

A kémiai elemekre vonatkozó mérési eredményeket rendre statisztikai sokaság elemeinek tekinthetjük. Jelen esetben az elemszámot, mely 1026 és 1100 volt, elég nagynak tartjuk ahhoz, hogy azt elméletileg végtelen elemű sokaságként tekintsük és felhasználjuk a tárgyalt statisztikai fogalmak szemléltetésére, anélkül, hogy a matematikai elmélet követelményei jelentősen sérülnének.

Azonban mivel a valóságban véges sokaságot kaptunk meghatározhatóvá váltak a valószínűségi változók várható értékei és szórásaik. Ezt az *I. táblázat* mutatja be.

A Ca, As és P változók sokaságaiból annak bemutatására, hogy a minta elemei valószínűségi változók, a sokaságból véletlenszerűen 100 elemű mintákat vettünk, 1000-szer. *A II. táblázat* Ca-ra vonatkozó minták realizációiból mutat be részleteket. Jól követhető, hogy a 100 elemű minták realizációi mintáról mintára változnak.

I. táblázat. A sokaság paraméterei
Table I The parameters of the manifold

	Mintaszám (db)	Átlag (mg/kg)	Szórás(mg/kg)
Ca	1026	71429,79	25301,74
As	1110	17,16	23,33
P	1026	1652,86	1014,58

II. táblázat. A minta realizációi
Table II Realisations of Ca-samples

	Minta realizáció						
Kalcium	X_1	X_2	X_3	X_4	...	X_{99}	X_{100}
1. minta	58339,44	52771,68	59438,88	44729,46	...	80643,21	73601,96
2. minta	78266,18	82664,06	59843,29	61782,66	...	55465,67	47424,37
.							
.							
.							
1000. minta	51623,78	59682,54	45447,23	50109,89	...	8806,80	8452,38

A *II. táblázat* mintáiból alapstatisztikák számíthatók, amelyek közül az egyik legfontosabb, a mintaátlag kerül bemutatásra a *III. táblázatban*. A táblázatot két további valószínűségi változóval bővítettük, az arzénnel és foszforral. A táblázat adatai szemléletesen látattják, azt az állítást, hogy a mintaátlag is valószínűségi változó, mintáról mintára változik, és értékei szóródnak a sokasági átlag (*I. táblázat*) körül.

III. táblázat A mintaátlag realizációi
Table III Realisations of the sample means

	Mintaátlag		
	As	Ca	P
1. minta	20,94	66420,70	1741,25
2. minta	15,77	69815,47	1775,85
.			
.			
.			
1000. minta	14,36	74512,84	1542,36

Természetesen az összes lehetséges mintaátlag átlaga adja a sokasági átlagot, azaz a várható étéket Ez a tulajdonság a becslés torzítatlanságát jelenti:

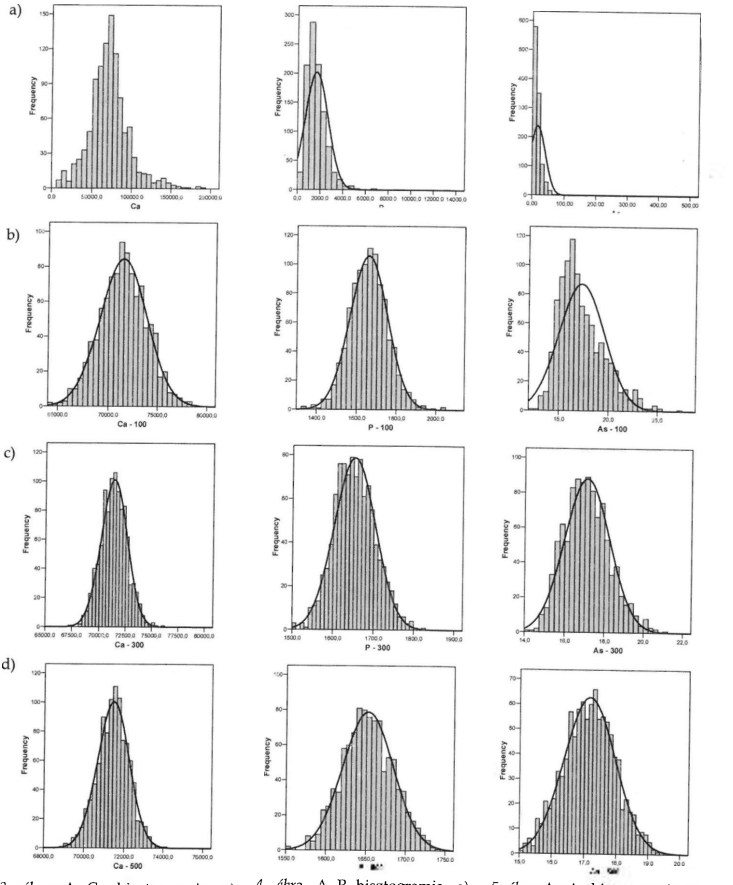
$$E(\overline{X}) = m$$

Ezt az elméleti megállapítást csak akkor lehetne bemutatni, ha az összes lehetséges mintaátlagot figyelembe vennénk. Ez azonban nehézségekbe ütközik, mivel például a Ca esetében az általunk sokaságnak tekintett 1026 mintaelemből, $9{,}51 \times 10^{140}$ módon lehet 100 elemű mintát kiválasztani. Más szavakkal: ennyi féle 100 elemű minta realizációt vagyunk képesek ebből a sokaságból előállítani és következésképpen ennyi különböző átlagot. (500 elemű minta kiválasztására $1{,}288 \times 10^{307}$ lehetőség van.) Ennek teljesítése gyakorlatilag lehetetlen. Ezért csak annak bemutatására lehet vállalkozni, hogy példánkon mutassuk be: a mintaátlag jól közelíti a sokasági átlagot, és hibája csökken ha a minta elemszáma elég nagy. Ezt a következőképpen valósítottuk meg. A sokaságokból 100, 300, 500 elemű mintákat vettünk, szintén 1000-szer. Kiszámítottuk a mintaátlagok átlagát és a rendre az átlagok hibáit. Az eredményekből néhányat a *IV. táblázat* tartalmaz. Az adatok a gyakorlatban is meggyőznek a fenti állításunkról.

IV. táblázat A mintaátlagok átlagai
Table IV Averages and standard errors of the sample means

Valószínűségi változó-mintarealizáció	Mitavételezés száma	Átlagok átlaga	Átlagok standard hibája	Átlagok szórása
Ca-100	1000	71498,90	74,72	2362,73
Ca-300	1000	71399,45	39,01	1233,64
Ca-500	1000	71444,89	25,12	794,45
As-100	1000	17,26	0,07	2,29
As-300	1000	17,12	0,04	1,13
As-500	1000	17,16	0,03	0,80
P-100	1000	1658,98	2,98	94,33
P-300	1000	1652,99	1,60	50,69
P-500	1000	1652,09	1,00	31,62

Vizsgáljuk meg a *IV. táblázatban* szereplő Ca, As, és P változók mintaátlagainak eloszlását. A kalciumnak, foszfornak és arzénnek mint sokaságnak (*3a, 4a, 5a ábrák*) és az ezekhez tartozó 100 (*3b, 4b, 5b ábrák*), 300 (*3c, 4c, 5c ábrák*) és 500 (*3d, 4d, 5d ábrák*) elemű minták átlagainak tapasztalati sűrűségfüggvényeit is megjelenítettük. Mind a három esetben látható, hogy a mintaátlag a minta elemszám-növekedésével – függetlenül attól, hogy az alapsokaság milyen eloszlású volt – normális eloszlást követ.



3. ábra. A Ca hisztogramja. a) „sokaság", b) 100, c) 300, d) 500 elemű minták átlaga

Fig. 3 The histograms of the complete Ca manifolds. The empirical probability density estimations of the sample averages for 1000–1000 subsamples, 100 (3.b), 300 (3.c) and 500 (3.d) long each Fig. 3

4. ábra. A P hisztogramja. a) „sokaság", b) 100, c) 300, d) 500 elemű minták átlaga

Fig. 4 a) The histograms of the complete P manifolds. The empirical probability density estimations of the sample averages for 1000–1000 subsamples, 100 (4.b), 300 (4.c) and 500 (4.d) long each

5. ábra. Az As hisztogramja. a) „sokaság", b) 100, c) 300, d) 500 elemű minták átlaga

Fig. 5 a) The histograms of the complete As manifolds. The mpirical probability density estimations of the sample averages for 1000–1000 subsamples, 100 (5.b), 300 (5.c) and 500 (5.d) long each

Összefoglalás

Egy környezetszennyezés elemzési eredményeinek felhasználásával próbáltuk bemutatni a mintát és annak néhány tulajdonságát. Definiáljuk a mintát úgy, hogy ez a meghatározás mind elméleti, mind gyakorlati szempontból kielégítő legyen. Ekkor a következő meghatározást javasoljuk:

A gyakorlati élet mintának nevezi valamely vizsgált jelenség adott paraméterének x, y, z, t koordinátákhoz, vagy azok intervallumához köthető, *in situ* mért, elemzett vagy az előbbiekből számított értékét. A gyakorlati értelemben vett minta, a matematikai minta egy elemének felel meg, azzal a különbséggel, hogy vonatkoztatási térfogata nagyobb, mint nulla.

Irodalom – References

ANDERSON, C. W. & LOYNES, R. M. 1987: The Teaching of Practical Statistics. – John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 212 p.

ANDÓ J. & BATA G. (eds.) 2001a: Budapest, XXII. Metallochemia telephelyet környező területek részletes tényfeltárási záródokumentációja. – Repét Kft. adattár, Budapest, Kézirat, 157 p.

ANDÓ J. & BATA G. (eds.) 2001b: Budapest, XXII. Metallochemia telephely környezeti állapotértékelő dokumentációja és a környezetvédelmi műszaki védelmi alternatívák vizsgálata. – Repét Kft. adattár, Budapest, Kézirat, 208 p.

BATA G. & ANDÓ, J. 2005: Nehézfémmel szennyezett talajszelvény környezeti minosító vizsgálata a Metallochemia-telephely (Budapest, XXII. kerület) környezetében. – *Földtani Közlöny* **135/1**, 91–111.

BATA, G., CSÁNYI, V. & KOVÁCS, J. (in press): Applied GIS System in Environmental Planning at Nagytétény Bay. – *Acta Geodaetica et Geophysica Hungarica.*

FÜST A. 1998: Geostatisztika. – Eötvös Kiadó, Budapest.

NEMETZ T. & WINTSCHE G. 1999: Valószínűségszámítás és statisztika mindenkinek. – Plygon, Szeged.

WEBSTER, R. & OLIVER, M. A. 1990: Statistical Methods in Soil and Land Resource Survey. – Oxford University Press, 568 p.

Kézirat beérkezett: 2005. 02. 15.